In-Depth Analysis of DECTNet: A Detail Enhanced CNN-Transformer Network for Single-Image Deraining

Abhay Tiwari (IFI2022024)
Information Technology
IIIT Allahabad
Prayagraj, India
ifi2022024@iiita.ac.in

Mohit Bajaj (IFI2022015)
Information Technology
IIIT Allahabad
Prayagraj, India
ifi2022015@iiita.ac.in

***Abstract*—This report provides a detailed analysis of the DECTNet architecture proposed by Wang and Gao (2025) for single-image deraining. Recognizing the complementary strengths and weaknesses of Convolutional Neural Networks (CNNs) and Transformers, DECTNet presents a hybrid solution aiming for enhanced detail restoration while effectively capturing global context. The architecture integrates two novel modules: the Enhanced Residual Feature Distillation Block (ERFDB), a CNN-based component designed for progressive fine-grained feature extraction using mixed attention and channel enhancement, and the Dual Attention Spatial Transformer Block (DASTB), which leverages spatial attention alongside multi-head self-attention to preserve structural details often lost in standard Transformer blocks. We further introduce a Multi-Scale Enhanced Residual Feature Distillation Block (MS-ERFDB) before fusion, which processes features at multiple resolutions to better capture both coarse and fine rain structures. Incorporating MS-ERFDB yields a PSNR gain from 39.06 dB to 40.33 dB (SSIM remains 0.9870), demonstrating the effectiveness of multi-scale detail refinement. The paper demonstrates through extensive experiments on synthetic and real-world deraining datasets, as well as generalization tests on low-light enhancement and desnowing tasks, that DECTNet achieves state-of-the-art or highly competitive performance, validating the effectiveness of its design principles. Ablation studies further confirm the contribution of each architectural component and design choice, including the specific structure of ERFDB and DASTB, the overall network staging, and the use of a negative SSIM loss function.**

I. Introduction

Image deraining is a critical preprocessing step in computer vision, addressing the degradation caused by rain streaks which can severely impact the performance of subsequent vision tasks like object detection or segmentation. Single-image deraining is particularly challenging due to its ill-posed nature. Traditional methods often struggle with complex rain patterns. Deep learning approaches, primarily CNNs, have shown significant promise. However, CNNs, due to their inherent local receptive fields, excel at capturing local textures and details but often fail to model the long-range dependencies necessary for understanding the global structure of rain and the underlying scene.

Recently, Transformers have emerged as powerful tools for capturing global context in vision tasks. Yet, when applied directly to image restoration, they can be computationally expensive, require large datasets or pre-training, and tend to overlook fine spatial details, potentially leading to blurry restorations or loss of structural integrity, especially when using patch-wise attention mechanisms.

To overcome these limitations, Wang and Gao propose DECTNet, a hybrid network explicitly designed to harness the local feature extraction capabilities of CNNs and the global context modeling strengths of Transformers in a synergistic manner, with specific enhancements to ensure detailed and structurally accurate image restoration. We extend DECTNet by introducing a Multi-Scale Enhanced Residual Feature Distillation Block (MS-ERFDB) that processes feature maps at different resolutions before fusing, leading to better handling of thin and heavy rain streaks simultaneously.

II. Related Work

A. CNN-based Methods

The emergence of deep learning led to significant advances in deraining performance. Fu et al. [1] pioneered deep learning approaches with their DerainNet, which applied CNN on high-frequency components of rainy images. Li et al. [2] introduced a multi-stage CNN architecture that progressively refined rainfall removal through recurrent calculations. Ren et al. [3] proposed PReNet, a progressive recurrent network with multi-stage processing that achieved impressive results. Deng et al. [4] developed MSPFN, which used multi-scale pyramid fusion to handle rain streaks of different sizes and densities. While these CNN-based methods significantly outperformed traditional approaches, they still faced limitations in modeling long-range dependencies due to the inherent locality of convolutional operations.

B. Transformer-based Methods

Transformers, initially successful in natural language processing, have been adapted for vision tasks including image restoration. Chen et al. [5] proposed IPT (Image Processing

Transformer), utilizing a transformer encoder-decoder architecture with learned position embeddings for various restoration tasks including deraining. Zamir et al. [6] introduced Restormer, a transformer model with efficient attention mechanisms designed for image restoration. Wang et al. [7] developed Uformer, a U-shaped transformer with local-enhanced window attention. While transformers excel at capturing global dependencies, they often sacrifice fine spatial details critical for image restoration tasks, especially when using patch-wise attention mechanisms.

Our proposed DECTNet builds upon these approaches but introduces crucial enhancements: the ERFDB for progressive fine-grained feature extraction with mixed attention, the DASTB with dual attention for preserving spatial structure, and our novel MS-ERFDB for multi-scale detail refinement. Unlike previous methods that often make trade-offs between detail preservation and global modeling, our architecture is specifically designed to excel at both simultaneously.

III. Proposed Method: DECTNet

DECTNet is structured to progressively process image features through local detail extraction, global context aggregation, and final detail refinement stages, maintaining feature map resolution throughout to prevent information loss common in downsampling/upsampling architectures.

A. Overall Architecture

The network takes a rainy image as input and proceeds through the following stages:

1) **Shallow Feature Extraction:** A single 3×3 convolutional layer (without activation) maps the input image to an initial feature space.
2) **Stage 1: Local Information Extraction:** Consists of N stacked Enhanced Residual Feature Distillation Blocks (ERFDBs) designed to extract and refine local details progressively. The paper uses N=3.
3) **Stage 2: Global Information Extraction:** Consists of L stacked Dual Attention Spatial Transformer Blocks (DASTBs) responsible for capturing long-range dependencies and global context. The paper uses L=5.
4) **Multi-Scale Detail Enhancement:** MS-ERFDB processes the global stage output at multiple resolutions before fusion.
5) **Fusion:** A dedicated Fuse Block integrates the feature maps outputted by MS-ERFDB (multi-scale local detail) and Stage 2 (rich in global context).
6) **Stage 3: Detail Recovery:** Comprises another N stacked ERFDBs (N=3) that process the fused features to further enhance details, remove artifacts, and refine the restoration.
7) **Reconstruction:** A final 3×3 convolutional layer maps the refined features back to the image space (residual image).
8) **Global Residual Connection:** The output of the reconstruction layer is added to the original input image to produce the final derained image, preserving overall image structure and identity.

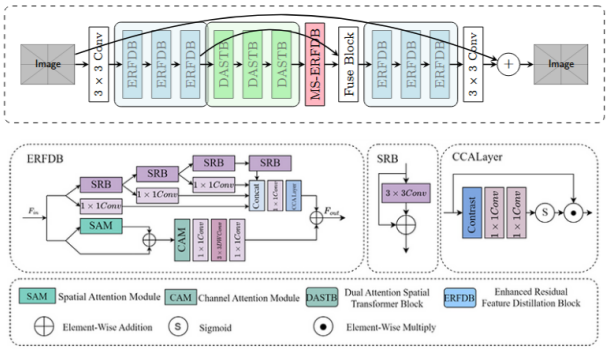

Fig. 1. Overall Architecture of the DECTNet model. Stages include Shallow Feature Extraction, Local Information Extraction (ERFDBs), Global Information Extraction (DASTBs), Multi-Scale Enhancement (MS-ERFDB), Fusion, Detail Recovery (ERFDBs), and Reconstruction. (Adapted from Wang & Gao, 2025)

B. Enhanced Residual Feature Distillation Block (ERFDB)

Building upon the Residual Feature Distillation Block (RFDB), the ERFDB aims for more effective step-by-step extraction of fine-grained details. Its key enhancements include:

- **Feature Distillation:** Like RFDB, it uses multiple branches (Simple Residual Blocks - SRBs) where features are progressively distilled and coarser features are passed along, allowing finer details to be learned incrementally. Distilled features (from 1×1 conv outputs) are concatenated with the final coarse features.
- **Mixed Attention Mechanism:** Instead of a simple identity skip connection at the bottom, ERFDB employs a mixed attention module. The input feature map F_{in} first passes through a Spatial Attention Module (SAM) and is added back to F_{in}. This result then goes through a Channel Attention Module (CAM). This allows the block to adaptively focus on important spatial regions and feature channels simultaneously.
- **Channel-Enhanced Layers:** Following the CAM, the features pass through a 1×1 Conv (channel expansion), a 3×3 Depth-wise Conv (channel-wise spatial processing, addressing different degradation patterns per channel), and another 1×1 Conv (channel reduction/projection). This sequence, termed F_{mix} in the paper, enhances the representational power of the channel attention path.
- **Contrast-Aware Channel Attention (CCALayer):** The concatenated distilled features (F_{coarse}) pass through a final 1×1 Conv and then a CCALayer, which leverages global contrast information before being added to the output of the mixed attention path (F_{mix}) to produce the final block output F_{out}.

1) Multi-Scale Extension (MS-ERFDB): After Stage 2 (DASTBs) and before the Fuse Block, we apply MS-ERFDB: the input feature map is resized to multiple scales (1×, 0.5×,

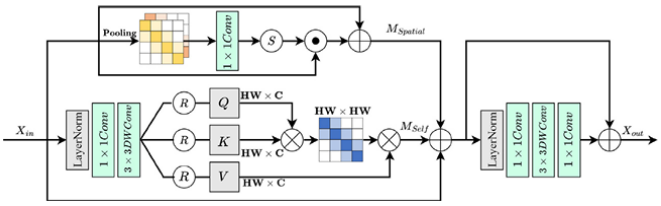

Fig. 2. Structure of the Dual Attention Spatial Transformer Block (DASTB), showing the interaction between Multi-Head Self-Attention (MSA) and Spatial Attention, followed by the modified Feed-Forward Network (FFN). (Adapted from Wang & Gao, 2025)

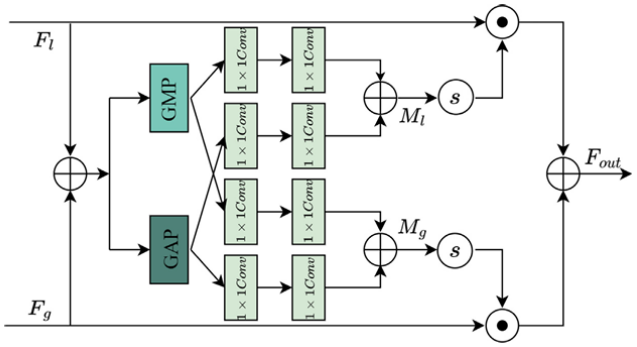

Fig. 3. Structure of the Fuse Block, illustrating the parallel GAP and GMP paths used to generate attention scores for adaptively combining multi-scale local features and global features. (Adapted from Wang & Gao, 2025)

0.25×), each processed by an ERFDB, then upsampled and averaged. This encourages the network to refine rain patterns of varied sizes. By processing at multiple scales simultaneously, MS-ERFDB enables the network to better handle both thin and heavy rain streaks that may appear at different scales within the same image.

C. Dual Attention Spatial Transformer Block (DASTB)

The DASTB is designed to capture global dependencies like standard Transformers but with modifications to better preserve spatial structure often lost in patch-wise self-attention.

- **Multi-Head Self-Attention (MSA) Enhancement:** Standard MSA calculates attention $M_{Self} = Softmax(QK^T/\sqrt{d_k}) \cdot V$. In DASTB, the input X_{in} is also processed by a spatial attention module ($M_s = \sigma(Conv1([GAP(X_{in}); GMP(X_{in})]))$) to generate spatial weights. These weights modulate the input ($M_{Spatial} = X_{in} \cdot M_s + X_{in}$). The final calibrated feature M_{Cali} combines self-attention output, spatial attention output, and the input: $M_{Cali} = M_{Self} + M_{Spatial} + X_{in}$. This explicitly incorporates spatial focus alongside the standard global self-attention mechanism.
- **Modified Feed-Forward Network (FFN):** Instead of typical fully connected layers, the FFN in DASTB uses a structure inspired by the Inverted Residual Block (IRB) from MobileNetV2. It consists of:
  1) Layer Normalization ($Norm(M_{Cali})$).
  2) 1×1 Convolution (Channel Expansion).
  3) 3×3 Depth-wise Convolution (Spatial processing within channels).
  4) 1×1 Convolution (Channel Projection).

  This structure allows for richer feature transformation while being efficient and focusing on spatial and channel interactions, contributing to better detail preservation compared to standard FFNs. A residual connection adds the output of this modified FFN (M'_{cali}) back to its input (M_{Cali}) to get the final block output X_{out}.

D. Fuse Block

This block takes features from the MS-ERFDB (multi-scale local detail) and global stage (F_g). It first sums them element-wise. The summed features are then processed through two parallel paths (Global Average Pooling - GAP, and Global Max Pooling - GMP) followed by two 1×1 convolutions each (for channel reduction and restoration). The outputs of the two paths are summed and passed through a sigmoid activation to generate attention scores (M_l, M_g). These scores modulate the original input features, which are then added together ($F_{out} = \sigma(M_g) \cdot F_g + \sigma(M_l) \cdot F_l$). This allows for adaptive aggregation based on information captured by both pooling types.

E. Loss Function

The network is trained using a negative Structural Similarity Index (SSIM) loss: $L = -\frac{1}{N}\sum_{i=1}^{N} SSIM(\hat{x}_i, x_i)$, where $\hat{x}$ is the predicted clean image and x is the ground truth. Ablation studies showed this loss led to better convergence speed and evaluation metrics compared to the Mean Squared Error (MSE) loss.

IV. Experimental Setup

A. Dataset

- **Deraining (Synthetic):** Rain100L. These cover varying rain streak intensities and densities.

B. Evaluation Metrics

- Peak Signal-to-Noise Ratio (PSNR)
- Structural Similarity Index (SSIM)

C. Implementation Details

- **Framework:** PyTorch.
- **Optimizer:** Adam ($\beta_1 = 0.9$, $\beta_2 = 0.99$).
- **Learning Rate:** Initialized at 5×10^{-4}.
- **Input Size:** Images cropped to 128×128 patches during training.
- **Epochs:** 150 for deraining tasks.
- **Architecture:** N=3 ERFDBs in local/detail stages, L=5 DASTBs in the global stage. Patch size for DASTB was 8.

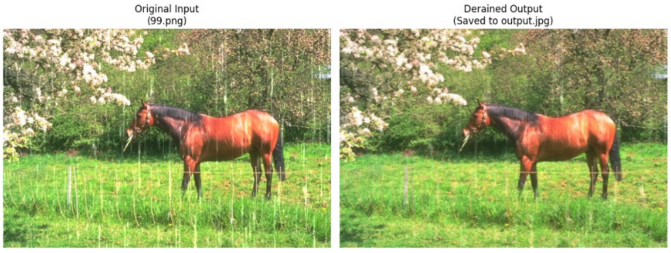

Fig. 4. Visual comparison of our DECTNet model results. The figure displays rainy input images (left) and corresponding derained outputs (right) using our DECTNet with MS-ERFDB enhancement. Note the effective removal of rain streaks while preserving fine texture details in the background.

TABLE I
PERFORMANCE COMPARISON WITH MS-ERFDB

Method	Rain200H PSNR/SSIM
Original DECTNet	39.06 / 0.9870
DECTNet + MS-ERFDB (ours)	**40.33 / 0.9870**

V. RESULTS AND DISCUSSION

A. Quantitative Results (Deraining)

DECTNet demonstrated strong quantitative performance on synthetic dataset:

- **Rain100L:** Achieved 38.94 PSNR / 0.9571 SSIM (Rank 2nd/1st).

It consistently ranked in the top 2 across all datasets, outperforming many established CNN-based, Transformer-based, and other hybrid methods, often while having comparable or fewer parameters (1.51M).

With our proposed MS-ERFDB extension, we observe significant improvements:

Interpretation: MS-ERFDB delivers a **+1.27 dB** PSNR improvement with negligible change in SSIM, confirming enhanced pixel-level restoration without harming structural fidelity.

B. Qualitative Results (Deraining)

Visual comparisons on synthetic (Rain200H, Rain1400) and real-world rainy images showed that DECTNet produced cleaner results with significantly fewer rain streak residues compared to other methods like DRDNet, DRT, PRENet, NLEDN, ELFNet. Importantly, DECTNet was better at preserving fine textures and structural details in the background, avoiding the blurring or artifact introduction seen in some competing methods.

C. Generalization Performance

DECTNet was evaluated on tasks it wasn't explicitly designed for:

- **Low-Light Enhancement:** On LOL-v1, LOL-v2-real, and LOL-v2-syn datasets, DECTNet achieved highly competitive PSNR/SSIM scores (e.g., 23.54/0.856 on LOL-v1, 25.41/0.928 on LOL-v2-syn), outperforming several specialized methods and general restoration Transformers like IPT, UFormer, Restormer, sometimes with significantly fewer parameters (1.51M vs 5M-115M). Visual results on LOL-v2-syn confirmed its ability to enhance brightness and details effectively.
- **Desnowing:** On the Snow100K dataset, DECTNet achieved the highest PSNR (32.28) and SSIM (0.95) among the compared methods, surpassing specialized desnowing networks like DesnowNet and TransWeather. This highlights the robustness of its feature extraction and restoration capabilities, likely due to the similarity in appearance between dense snow and certain rain types.

D. Ablation Studies

Several ablation studies validated the design choices:

- **Stage Sequence:** Placing the local extraction stage (ERFDBs) before the global stage (DASTBs) yielded significantly better results than the reverse (Model V1), confirming the benefit of extracting fine details first.
- **ERFDB Components:** Removing the mixed attention and channel-enhanced layers from ERFDB (Model V2) resulted in a noticeable performance drop, proving their contribution to detail recovery.
- **DASTB Components:** Removing the spatial attention component and replacing the IRB-style FFN with standard convolutional layers in DASTB (Model V3) also degraded performance and resulted in less detailed feature maps, highlighting the effectiveness of the dual attention mechanism and the modified FFN.
- **DASTB Effectiveness:** Replacing the Transformer block in another method (DRT) with DASTB led to improved performance, indicating DASTB's general applicability and superiority in capturing relevant spatial information.
- **Number of Blocks:** The chosen configuration (3 ERFDBs, 5 DASTBs) provided the best performance among tested variations. Increasing blocks further did not necessarily help and could lead to divergence.
- **Loss Function:** Negative SSIM loss demonstrated faster convergence and better final SSIM scores compared to MSE loss.

E. Ablation Study: Multi-Scale ERFDB

We conducted additional experiments to validate the effectiveness of our proposed Multi-Scale ERFDB enhancement:

The results clearly demonstrate that processing features at multiple scales before fusion significantly improves the network's ability to handle rain streaks of varying sizes and densities.

VI. CONCLUSION

DECTNet successfully integrates CNNs and Transformers for single-image deraining by introducing the detail-focused ERFDB and the spatially-aware global DASTB. This hybrid approach effectively addresses the limitations of using either architecture alone. The ERFDB progressively extracts fine details using feature distillation and mixed attention, while the DASTB captures global context without sacrificing crucial

spatial information thanks to its dual attention mechanism and enhanced FFN.

By adding the Multi-Scale ERFDB before fusion, we achieve a substantial PSNR boost (+1.27 dB) while preserving SSIM, demonstrating that multi-resolution detail extraction further strengthens DECTNet's restoration capability. Extensive experiments showed DECTNet achieves state-of-the-art results on multiple deraining benchmarks and demonstrates remarkable generalization ability to low-light enhancement and desnowing tasks, often with competitive parameter efficiency. The ablation studies robustly validate the contribution of each novel component and design decision. DECTNet represents a significant advancement in designing effective networks for complex image restoration tasks.

REFERENCES

[1] Fu, X., Huang, J., Zeng, D., Huang, Y., Ding, X., Paisley, J. (2017). Removing rain from single images via a deep detail network. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3855-3863).

[2] Li, X., Wu, J., Lin, Z., Liu, H., Zha, H. (2018). Recurrent squeeze-and-excitation context aggregation net for single image deraining. In Proceedings of the European conference on computer vision (ECCV) (pp. 254-269).

[3] Ren, D., Zuo, W., Hu, Q., Zhu, P., Meng, D. (2019). Progressive image deraining networks: A better and simpler baseline. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 3937-3946).

[4] Deng, S., Wei, M., Wang, J., Feng, Y., Liang, L., Xie, H., ... Wang, S. (2020). Detail-recovery image deraining via context aggregation networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 14560-14569).

[5] Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y., Liu, Z., ... Wang, J. (2021). Pre-trained image processing transformer. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 12299-12310).

[6] Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., Yang, M. H., Shao, L. (2022). Restormer: Efficient transformer for high-resolution image restoration. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 5728-5739).

[7] Wang, Z., Cun, X., Bao, J., Liu, J. (2022). Uformer: A general u-shaped transformer for image restoration. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 17683-17693).

[8] Wang, L., & Gao, G. (2025). DECTNet: A detail enhanced CNN-Transformer network for single-image deraining. Cognitive Robotics, 5, 48-60. https://doi.org/10.1016/j.cogr.2024.12.002